Haplotype Function Score improves biological interpretation and cross-ancestry polygenic prediction of human complex traits

Weichen Song[1,2]✉, Yongyong Shi[2,3]✉ and Guan Ning Lin[1]✉

1 Shanghai Mental Health Center, Shanghai Jiaotong University School of Medicine, School of Bioengineering, Shanghai Jiaotong University, Shanghai, China
2 Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders (Ministry of Education), Collaborative Innovation Center for Brain Science, Shanghai Jiao Tong University, Shanghai, China
3 Biomedical Sciences Institute of Qingdao University (Qingdao Branch of SJTU Bio-X Institutes), Qingdao University, Qingdao, China

✉Correspondence:
Dr. Weichen Song, goubegou@sjtu.edu.cn
Prof. Yongyong Shi, shiyongyong@gmail.com
Prof. Guan Ning Lin, nickgnlin@sjtu.edu.cn

Haplotype Function Score improves biological interpretation and cross-ancestry polygenic prediction of complex traits

Abstract

We proposed a new framework for human genetic association study: at each locus, use deep learning model (in this study, Sei) to calculate functional genomic activity score for two haplotypes per individual, and use this score, defined as haplotype function score (HFS), to replace original genotype in association study. Applying HFS framework to 14 complex traits in UK Biobank, we identified 3,619 independent HFS-trait association with $p<5\times10^{-8}$. Fine-mapping revealed 2,699 causal associations, corresponded to a median of 63 more causal findings per trait compared with SNP-based analysis. HFS-based enrichment analysis revealed 727 pathway-trait association and 153 tissue-trait association with strong biological interpretability, including "circadian pathway-chronotype" and "Arachnoid acid-intelligence". Lastly, we trained neural network and applied transfer learning to integrate with SNP-based polygenic risk score, which showed 4.1% to 28.8% improvement in the cross-ancestry polygenic prediction. We concluded that HFS is a promising strategy for understanding genetic basis of human complex traits.

Introduction

Genome-wide association studies (GWAS) have witnessed remarkable advancements over recent years, both in terms of sample size and genetic discovery. However, the elucidation of downstream mechanisms and subsequent applications still face certain limitations[1]. One caveat is that the statistical power of GWAS on a variant relies on its population frequency[2–4], whereas most variants with large effect size are rare[5], leading to insufficient discoveries. Moreover, Linkage Disequilibrium (LD) among neighboring variants can significantly inflate false positive results[6]. The variability of LD structure among different populations further compounds the challenges associated with training predictive models and discovering causal genes. Lastly, most trait-relevant variants reside in non-coding regions[7], which lack direct functional annotations as coding variants. The prevalent approach to addressing this issue is to annotate each variant based on its location within functionally significant regions[8–12], such as transcription factor binding sites or enhancers. While this strategy has considerably advanced the analysis, it's not optimal, as a variant's placement within a functionally important region does not inherently signify that the variant has substantial functional impacts.

The central dogma, proposing that DNA alterations' effects on phenotype are mediated via RNA and protein changes, offers a novel strategy to address these challenges. More precisely, by replacing the original genotypes in association studies

with the aggregated impact of variants on transcription or functional genomics, the central dogma ensures the preservation of the majority of genetic information. This 'aggregated impact' offers several benefits for GWAS analysis: it provides direct biological interpretations, bypasses the effects of LD and population genetic history, and amalgamates information from both common and rare variants. One successful implementation of this strategy is Polygenic Transcriptome Risk Scores (PTRS)[13,14], which employ genetically determined transcription levels rather than genotypes to predict complex trait, and achieved remarkable portability. Nonetheless, the accuracy of imputing transcription levels from genotypes, given the sample size of currently available cohorts (such as the Genotype-Tissue Expression project, GTEx[15]), remains limited (R2 around 0.1 for most genes)[16]. Thus, the performance of PTRS is yet to reach its optimal potential.

Following the success of PTRS, we made one step forward to utilize functional genomics in this strategy. Compared with transcription levels, predicting genetically-determined functional genomic levels has achieved much higher accuracy by multiple recent deep learning (DL) studies[17–21]. These DL models utilize segments of the human reference genome as training samples, substantially increasing the sample size. Furthermore, functional genomics serve as a mediator between DNA and transcription, thus lessening the influence of non-genic factors such as the environment. Given these advancements, we propose that using the outputs of one of the state-of-the-art DL models, Sei[17], as the 'aggregated impact' in this novel strategy could effectively address the challenges aforementioned. Sei accepts a DNA sequence and computes multiple sequence class scores that represent different facets of the functional genomic activities of that sequence. This score integrates impacts from all variants, even those as rare as singletons, into one continuous variable, and is, in theory, unaffected by LD.

In this study, we present an analytical framework founded on this strategy (Figure 1) and implement it on complex traits in the UK Biobank to pinpoint causal loci and genes, decipher biological mechanisms, and devise cross-ancestry prediction models. We segmented the human reference genome into multiple 4,096 bp loci, generated DNA sequences for each locus for two haplotypes per individual, and employed Sei to compute the functional genomic activities of these sequences. We designated this activity score as the Haplotype Function Score (HFS) and analyzed the association between the HFS and each trait. Our findings confirm that the HFS framework offers a unique improvement in the biological interpretation and polygenic prediction of complex traits compared to classic SNP-based methods, thereby demonstrating its value in genetic association studies.

Result

Overview of genome-wide HFS

We used the HFS framework to analyze imputed genotype data from the UK Biobank (Figure 1). We segmented the human genome (hg38) into 617,378 discrete, non-overlapping loci, each 4,096 base pairs long. Of these, 590,959 loci carried at least one non-reference haplotypes in the UKB cohort (see Method and Table S1). After

quality control, these loci contained approximately 1.2 billion haplotypes, with a median count of 819 per loci (Figure S1). We then employed the deep learning framework, sei[17], to compute sequence class scores for each haplotype. In its sequence-mode, sei accepts DNA sequences in fasta format and produces multiple distinct sequence class scores, 39 of which were included in our study (Method). Our analysis identified significant variation in sequence class scores across different loci. In fact, 49.7% of loci housed haplotypes whose sequence class (as defined by the maximum of the 39 sequence class scores) differed from the reference haplotype sequence class. Using the reference sequence class as a benchmark, we noted that 16.8% of loci showed a difference between the maximum and minimum haplotype scores that surpassed the score of the reference haplotype. Moreover, the correlation between sequence class scores of adjacent loci was low, with a median R^2 value of 0.02 (Figure S2), effectively reducing the impact of LD in association studies.

Expanding on the sequence class scores, we obtained trait-specific HFS. For each locus, we computed the mean sequence class score of two haplotypes per individual, reflecting an additive model, as well as the minimum and maximum scores, indicative of a dominance model. We employed SUSIE[22] to select one of 117 possible scores (39 × 3) that was most likely associated with the trait under study. This selected score was designated as the HFS of that locus, and its association with the corresponding trait was computed using a generalized linear model. We applied this framework to 14 polygenic traits in the UKB British ancestry training set (n=350,587; Table S2 and Method), identifying 16,597 significant HFS-trait associations at a threshold of $p<5\times10^{-8}$ (n=15 for insomnia, n=7,573 for height; Table S2), equating to roughly 3,619 independent associations. The most significant associations were between the "promotor" score of chr7:121327898-121331994 and bone mineral density (BMD; regression beta=-0.02, $p<10^{-300}$), and the "promotor" score of chr9:4760952-4765048 and platelet count (beta=3.20, $p=2.79\times10^{-262}$; Table S3). The genomic control inflation factor (λ_{GC}) for the HFS association test varied between 0.99 for asthma and 1.50 for height, closely resembling the standard SNP-based GWAS on the same data (Pearson Correlation Coefficient [PCC]=0.91, paired t-test p=0.16; Method and Figure S3). These findings suggest that HFS-based association tests had adequate power and do not introduce additional p-value inflation.

Fine-mapping based on HFS

Based on these data, we applied SUSIE to fine-map the causal loci that were associated with each of the 14 traits. We divided hg38 genome into 1,361 independent blocks as defined by MacDonald et al.[23], and applied SUSIE to loci HFS in each of these blocks (number of loci per block=4~2392). As shown in Figure 2 and Table S4, we identified a total of 2,699 causal loci-trait associations at the threshold of posterior inclusion probability (PIP)>0.95, hereafter referred to as “causal loci”. Compared with SNP-based functionally-aware fine-mapping methods (PolyFun[9] and SbayesRC[11]), HFS-based SUSIE detected -11 to 334 more causal signals

(median=63, Table S5) for each trait. We cautioned that these methods use summary statistics as input and are by nature less sensitive than individual data-based method. Yet, we suggested that such impact would be mild, since we used in-sample LD reference (from UKB European sample). Among these causal loci, only 22% were also lead loci in association analysis (loci with the lowest p-value in 200kb region), and 58% had association p-value > 5×10^{-8}. In line with previous SNP-based analysis[9], this result highlighted the importance of using causal signals instead of lead signals in post-GWAS analysis. We found 67 causal loci showing pleiotropic effects on at least two independent traits, including "CTCF-Cohesin" score of chr9:89596537-89600633 that was associated with age at menarche, body mass index (BMI) and height (PIP>0.97; Table S4).

When looking at the reference sequence class of loci, those with functionally importance were more likely to be causal loci, including "Promoter" (Odds Ratio [OR]=2.33, $p=1.41\times10^{-14}$) and "Bivalent stem cell enhancer" (OR=2.22, $p=1.11\times10^{-8}$) and "Transcribed region 1" (OR=1.71, $p=1.581\times10^{-10}$, Figure 2). Such functional enrichment was even higher for pleiotropic loci ("Promoter": OR=7.20, $p=3.35\times10^{-5}$). We also observed trait-specific patterns of such sequence class enrichment, such as "CEBPB binding site" (Insomnia: OR=5.25, p=0.01) and "FOXA1/AR/ESR1 binding site" (intelligence: OR=4.69, p=0.01, Figure 2 and Table S6). These results demonstrated the expected functional patterns of causal loci, and indicated that HFS-based fine-mapping was biologically interpretable and reliable.

Despite the functional enrichment, we applied several secondary analyses to verify the reliability of HFS-based SUSIE result. Firstly, we took causal SNP fine-mapped by PolyFun[9] as positive control, and find that compared with genomic region-matched control loci, causal loci were significantly enriched for causal SNP (OR=1.33 to 5.08, Fisher test p=0.12 to 4.72×10^{-52}, Table S5). Secondly, we calculated the heritability tagged by causal loci and PolyFun causal SNP in independent test set (defined as the R^2 of linear regression; Method), and found that causal loci tagged 38% to 251% more heritability than causal SNP (median=151%; Table S5). This was not an artifact of larger number of causal loci, since the Akaike information criterion (AIC) was similar between causal loci and causal SNP (paired t-test p=0.36; Table S5). Thirdly, for traits with sufficient causal loci coverage, we also applied Linkage Disequilibrium Score regression (LDSC) on independent GWAS summary statistic to evaluate heritability enrichment in causal loci. On average, causal loci showed 124-fold enrichment of heritability, significantly larger than genomic region-matched control loci (124-fold vs 101-fold; p=0.0002, Method and Figure S4). Lastly, we applied simulation analysis and found that HFS+SUSIE showed similar advantages over SNP-based methods as in real data, with high accuracy and low false positive rate (Supplementary materials). Taken together, HFS-

based SUSIE is a powerful and robust strategy for individual data-based genetic fine-mapping.

Biological interpretation based on HFS

Pinpointing causal loci of complex traits provides the opportunity of analyzing the biological mechanism of them. Thus, based on the HFS-based fine-mapping result, we applied a linear regression model to analyze the underlying pathways, cell types and tissues of each complex trait. For each locus, we annotated its relevance to a pathway by combined SNP to Gene (CS2G) strategy[24], and regressed the PIP against this annotation on conditioned of a set of baseline annotations, similar to the LDSC framework[25] (Method). After p-value correction and recurrent pathway removal (Method), we detected a total of 727 pathway-trait association (Figure 3A and Table S7). The most significant associations were "megakaryocyte differentiation" with platelet count ($p=2.26\times10^{-34}$), "Insulin-like growth factor receptor signaling pathway", "Endochondral ossification" with height ($p=4.95\times10^{-33}$ and 1.17×10^{-27}), "PD-1 signaling" with allergic disease ($p=5.55\times10^{-25}$), and "major histocompatibility complex pathway" with asthma ($p=1.22\times10^{-23}$). In fact, asthma and allergic disease were predominantly associated with more than 80 immune-related pathways. These association were all in line with existing knowledge of trait mechanism, and extended the understanding of their genetic basis. For example, PD-1 has recently been suggested as potential targets of allergic diseases like atopic dermatitis[26], but such association has not been highlighted by previous genetic association studies.

For other traits, the most significant associations also replicated known mechanisms, such as "osteoblast differentiation", "Wnt ligand biogenesis and trafficking" with BMD ($p=4.59\times10^{-13}$ and 2.78×10^{-12}); "circadian pathway" with chronotype ($p=4.25\times10^{-12}$); "calcium regulated exocytosis of neurotransmitter", "Arachidonic acid metabolism" with intelligence ($p=5.52\times10^{-7}$ and 2.78×10^{-6}); "GPCR pathway" and "adipogenesis" with BMI ($p=4.97\times10^{-10}$ and 2.02×10^{-7}) and "physiological cardiac muscle hypertrophy" with systolic blood pressure ($p=6.32\times10^{-11}$). We also highlighted less significant association which provided novel insights, such as "synaptic vesicle docking" and "neuron migration" with chronotype ($p=4.00\times10^{-7}$ and 4.55×10^{-7}), "Prostaglandins synthesis" with insomnia ($p=5.30\times10^{-9}$), "behavioral response to cocaine" with alcohol intake ($p=3.39\times10^{-8}$) and "roof of

mouth development" and "glycoside metabolism" with FVC ($p=2.19\times10^{-12}$ and 5.73×10^{-11}).

For cell type and tissue analysis (Figure 3B and Table S8), we applied the same linear model to evaluate whether causal loci enriched in active chromatin regions of each cell type (Method). We found 153 biologically interpretable associations with complex traits. For example, fetal megakaryocyte ($p=5.67\times10^{-22}$) and child spleen ($p=2.15\times10^{-13}$) were found to be key cell type and tissue of platelet count. Systolic blood pressure was significantly associated with multiple heart and artery tissues and fetal cardiomyocyte ($p<1.63\times10^{-5}$), whereas allergic disease was associated with multiple immune cells including natural killer, Treg and B cells ($p<4.79\times10^{-16}$). For brain-related traits, we found 21 significant associations, 14 of which were from central nervous system. For example, adult Hippocampus and cingulate gyrus were both linked to alcohol intake, smoking and insomnia ($p<1.11\times10^{-5}$), whereas chronotype was associated with embryonic brain germinal matrix ($p<8.68\times10^{-6}$) and intelligence with embryonic neuron derived stem cell ($p<6.89\times10^{-7}$).

We also applied other modified strategies for this task but did not get satisfying result. For example, using cS2G to link locus to gene lists specifically expressed in each cell type suffered from scRNA dataset batch effect, whereas linear mix model was less sensitive than standard linear model (Supplementary Materials).

Taken together, our result suggests that fine-mapping result based on HFS could pinpoint the causal pathways, cell types and tissues underlying complex traits, and is valuable for the biological interpretation of genetic association study.

Highlighted genes for complex traits

Enhanced power of fine-mapping and biological enrichment could reveal novel key genes for trait mechanism study. Below we integrated fine-mapping result and their functional annotation in several case studies to find trait-relevant genes that were not highlighted by previous genetic association studies.

In our study, platelet count had large number of causal loci (Figure 2) which showed significant functional enrichment (Figure 3). To find key loci and genes underlying platelet count, we focused on causal loci that overlapped with active regions in "fetal megakaryocyte" and "child spleen tissue", and applied cS2G[24] to link them to two key pathways ("megakaryocyte differentiation" and "platelet morphogenesis", Method and Figure 4A). We chose these annotations based on p-value in biological enrichment analysis in Figure 3. A total of 25 loci were highlighted (Figure 4A), which were recurrently linked to well-known platelet-regulating genes like MEF2C, SH2B3, FLI1, RUNKX1, THPO and NFE2. Among them we noticed a

less-studied gene RBBP5, a target of key transcriptome factor MEF2C during megakaryopoiesis[27]. Specifically, in 1q32.1 region, HFS+SUSIE identified two loci with PIP>0.9 (Figure 4B), compared with SNP-based fine-mapping[9] that only found seven signals between PIP=0.1 to 0.5. This was unlikely a statistical inflation, since HFS-based association test p-value was actually higher than SNP-based one (Figure S5). One of the causal loci, chr1: 47401806-47405902 (PIP=1), overlapped with spleen active chromatin and harbored a cCRE in megakaryocyte, and was linked to RBBP5 and three other genes (Table S9). RBBP5 is known to involved in megakaryocyte differentiation during megakaryopoiesis and was regulated by MEF2C[27], but previous genetic association studies provided little evidence for its association with platelet count.

The major histocompatibility complex (MHC) region has long been a challenge of genetic association study due to its long-range LD, and is often excluded in fine-mapping tools. However, many disorders like Schizophrenia[28] and immune diseases[29] are robustly associated with MHC region. In our HFS-based fine-mapping of asthma, we found 15 loci within MHC region had PIP>0.95, 11 of which overlapped with active chromatin regions in Treg or natural killer cells (Figure 4C and Table S10). This result showed good discrimination between causal and non-causal loci: despite these 15 likely causal loci, only six loci had PIP between 0.25 to 0.95. Since MHC region harbored a large number of genes, these causal loci were linked to as much as 105 potential target genes, which hindered the discovery of true targets. We further filtered them based on the involvement in pathway "TNFR2-NFKB pathway" and "innate lymphocyte [ILC] development", since these pathways were most significantly associated with asthma (Figure 3), even after excluding MHC region ($p=2.57\times10^{-13}$ and 1.39×10^{-17}). We found five genes (LTA, LTB, TNF, PSMB8 and PSMB9) that were predicted to be regulated by five causal loci overlapped with active chromatin regions (Figure 4C), which could be considered as potential key genes for further validation.

Similarly, we fine-mapped MHC region for other allergic diseases (Figure S6 and Table S11) and found potential key genes including HLA family and AGER. We also highlighted other gene-trait association not previously emphasized by GWAS, including GATA4 and NPPA (cardiac muscle hypertrophy) with SBP, ALOX5 (Arachidonic acid metabolism) with intelligence and CRY1 (circadian pathway) with chronotype, as further discussed in Table S12 to S14 and supplementary information.

HFS-based polygenic prediction

Lastly, we analyzed the potentiality of HFS in polygenic prediction accuracy. Compared with state-of-the-art SNP-based polygenic risk score (PRS) algorithm LDAK-BOLT[30], HFS-based PRS (weighted by SUSIE posterior effect size) reached 43% to 90% of R^2 in independent European test set (meta-analyzed proportion= 74.5%, 95% confidence interval=74.3%~74.7%, Figure S7). The gap between performance of HFS-based and SNP-based PRS reflected the fact that HFS only captured (the majority of) functional genomic alterations and missed the information of amino acid sequence and post-translational modification. We thus proposed that

integrating information from HFS and SNP could provide better performance. Specifically, in the large European training set we trained SNP PRS model by LDAK and trained a fully-connected neural network (FCNN) based on HFS. Then, in a small tuning sample of target ancestry, we calculated the output of the last hidden layer of HFS-based FCNN as well as SNP PRS, and used elastic net to integrate them into a final polygenic prediction score, hereafter referred to as "HFS_{FCNN}+LDAK". This procedure resembled the transfer learning of FCNN (train the model in large sample, fix first n-1 layer, tune the output layer in small sample), and properly fitted the challenge that non-European training set for polygenic prediction is limited.

Using height as a representative trait, we first estimated the proportion of variance captured by top loci, and found that HFS of loci with PIP>0.4 (n=2,429) captured 80% of variance explained by all genome-wide loci (n=590,959; Figure 5A). Compared with other PIP threshold (Figure 5A), using PIP>0.4 achieved balance between performance and feature number, and is proper for FCNN given training sample size=350,587. After parameter tuning (Methods), we trained a FCNN on these 2,429 HFS with linear activation, two hidden layers and 133 nodes per layer, which showed 4.1% improvement in R^2 compared with HFS+SUSIE (Figure 5B). Using non-linear activation or more hidden layers and nodes worsen the prediction performance (Figure S8). This was in line with previous result that linear model missed neglectable heritability of height[31,32]. We then calculated HFS_{FCNN}+LDAK in non-British European (NBE), South Asian (SAS), East Asian (EAS) and African (AFR) population in UK Biobank, and observed 4.5%, 5.1%, 6.2% and 28.8% improvement over LDAK along ($p=1.15\times10^{-14}$, 0.0008, 0.004 and 0.004, respectively. Figure 5C). As a comparison, we integrated LDAK with PolyFun-pred[33] and SbayesRC[11] using Polypred framework[33], but did not observe significant improvement over LDAK along (difference in $R^2<0.01$, p=0.001~0.07, Figure 5C). Since PolyFun-pred+BOLT-LMM has been shown to significantly outperformed BOLT-LMM along[33], we reasoned that the improvement of LDAK over BOLT-LMM might have attenuated the improvement brought about by PolyFun-pred, making it difficult to reach significance threshold. Taken together, we concluded that HFS could bring about mild but significant improvement to classic SNP-based PRS in the task of cross-ancestry polygenic prediction.

Discussion

In this study, we designed the new HFS framework for genetic association analysis and demonstrated that it could improve classic SNP-based analysis in terms of causal loci and gene identification, biological interpretation and polygenic prediction. We suggest that HFS is a promising strategy for future genetic studies, but more progresses in algorithm and computation and data resources is still desired.

Compared with SNP, HFS has several compelling features. For instance, LD between adjacent HFS is much lower than SNP, which enhances the precision of statistical fine-mapping. For those false-positive variants caused by LD, they are expected to make little impacts on functional genomics, thus their HFS would be

close to reference and would not influence downstream analysis significantly. In line with these advantages, we showed that HFS-based fine-mapping had high statistical power, and downstream enrichment analysis was capable of revealing biologically interpretable mechanisms. As a typical example, our findings of enrichment of intelligence-associated loci in arachidonic acid metabolism pathway is in line with the well-known role of polyunsaturated fatty acid in neurodevelopment[34]. Nonetheless, previous GWAS provided little evidence on this association. Secondly, HFS could integrate effects of all variants within a locus, regardless of their population frequency. Thus, HFS could capture information from rare variants overlooked by classic association study and improve polygenic prediction, as shown by our result. In fact, HFS framework could directly extend to whole-genomes sequencing data and capture all mutations as rare as singleton, making one step forward to fill in the “missing heritability”.

Despite its potential, the current HFS framework carries several drawbacks and necessitates significant enhancements. A key limitation is the substantial computational cost. In this study, the transformation phase of the genotype-haplotype-sequence for UK Biobank SNP data required hundreds of thousands of CPU core hours. This computation cost would increase exponentially when analyzing whole-genome sequencing (WGS) data or employing a sliding-window strategy. A potential solution could involve developing a new algorithm that bypasses the variant calling stage and directly generates DNA sequences per locus from raw sequencing or SNP array data. For the sequence-to-HFS step, Sei[17] required about 1.8 GPU hours per one million sequences. Intriguingly, the majority of Sei's output is unused in the HFS framework, since Sei predicts over twenty thousand functional genomic features, while the HFS only represents one of their integrated scores. Future development of novel deep learning models that predict functional genomics in a manner more fitting to the HFS framework could considerably reduce computation costs. Lastly, it's currently unfeasible to incorporate all six hundred thousand genome-wide HFS into a single FCNN model. This limitation forced us to include only the top two thousand, sacrificing 20% of the accuracy.

Another hurdle arises in integrating HFS with other genomic features. Intrinsically, HFS captures only the variant effect mediated by functional genomics, while a genetic variant might also influence amino acids, post-transcriptional modifications (PTMs)[35], and 3D chromosomal structures[36]. Therefore, HFS alone cannot wholly replace SNP without any loss, as our results demonstrate that the HFS-based prediction model captured approximately 70% of the variance explainable by the SNP-based prediction model. One potential solution is to extend the concept of HFS, applying deep learning to quantify the genetically determined values of PTMs, protein biochemical properties[37], and protein and chromosomal structures, potentially employing AlphaFold[38]-derived features[39]. Analyzing HFS in conjunction with these multi-modal function scores could provide a comprehensive depiction of the genetic architecture of complex traits. However, the colossal computational cost is currently prohibitive. As a compromise, we simply performed joint analysis of HFS with SNP PRS in our prediction model analysis. This approach

is far from optimal, as it led to only moderate improvement and did not enhance fine-mapping and biological enrichment analysis.

In summary, our results demonstrate that incorporating HFS to represent genetically-determined functional genomic activities in genetic association studies offers robust improvements in both the biological interpretation and polygenic prediction of complex traits. Thus, the application of the HFS framework in future genetic association studies holds considerable promise.

Method

Sample description

This study analyzed UK Biobank data, with application ID 84436. We only included participants with array imputed genotype data in bgen format that passed UKB quality control, and removed related individuals. We randomly selected 350,587 British ancestry Caucasians as training sample. The remaining participants were grouped according to their ancestry, where Non-British European, South Asian, East Asian and African groups serve as test samples.

All phenotypes analyzed (Table S1) were collected from UKB table browser, which came from self-report or physical measurement. Phenotypes were first adjusted by age, sex, top ten principal components, Townsend index and genotype array quality metrics by linear regression. We then applied inverse-normal transformation on the residuals. Binary phenotypes were adjusted in the same way except by generalized linear regression.

Genotype data processing

We first segmented hg38 genome into 4,096 bp loci. To do so, we downloaded chromatin state annotation of 222 human tissues at different developmental stage (embryo, newborn, adult) from epimap[40] database. For each tissue, all chromosomal regions annotated as "transcription start site (TSS), transcription region (TX), enhancer, promoter" in at least half of the samples were marked as active regions. The union of active regions across all tissues was taken, and regions annotated as genomic gaps (centromere, ambiguous base pairs, etc.) in the Hg38 genome were removed. Then, for this series of active regions, if the length is less than 4096bp, the locus is defined as a 4096bp area centered around the active region. If the length is greater than 4096bp, 4096bp length loci are gradually delineated from the midpoint outward. Finally, non-overlapping 4096bp blocks were used to cover the remaining genomic regions. This resulted in about 617,378 genomic regions in total.

For each of the loci, we obtained ID of variants within this locus by bedtools[41], then extracted genotypes from UKB .bgen file by bgenix, finally used Plink[42] to remove all variants with INFO<0.8, Hardy-Weinberg $p<10^{-6}$, allele count<10 or missing rate >10%, and removed individual that missed more than 10% of retained variants in this locus. The output vcf file was liftover to hg38 by Crossmap[43] and phased by SHAPEIT4[44]. Phased vcf was transformed to .haps format by Plink, which in turn gave rise to two files: a vcf file containing information of each haplotype, and an n x 2 matrix in plain text that recorded the id of two haplotypes per individual.

HFS calculation

There has been several DL models that predict functional genomic profiles based on DNA sequence[17–21]. Among them, we chose sei[17] to calculate HFS for the following reasons: 1) the required input length (4,096bp) is moderate. 2) it represents 21,906 functional genomic tracks, more comprehensive than other models. 3) it represents overall activity of the entire sequence, not only the few bp at the center. For each haplotype at each locus, we generated its corresponding DNA sequence by bcftools[45] consensus option. At each locus, the start point of each sequence was matched to the start point of reference sequence. When insertion variants made the sequence longer than 4,096bp, we discarded base pairs at the 3' end. Likewise, with deletion variants, we added N to the 3' end. We applied sei to predict 21,906 functional genomic tracks for each sequence, without normalizing for histone mark. We then used the projection matrix provided by sei to calculate forty sequence class scores, which could be regards as the weighted sum of these tracks and represented different aspect of functional genomic activities. We discarded the last score (heterochromatin 6 [centromere]), since its proportion is too low and is functionally trivial, leading to 39 scores per haplotype.

On each individual, we derived three values from each sequence class score: mean of two haplotypes, corresponding to additive model; and maximum and minimum of two haplotypes, corresponding to dominance model. For HFS Linkage Disequilibrium calculation, we extracted the mean value of sequence class score corresponding to reference sequence class of adjacent loci, and calculate R^2 value between them, For trait-specific HFS calculation, we used the resulting 117 values as input to SUSIE[22] (parameter: coverage=0.95, maximum number of causal signal=1) and selected one value that was most probably associated with this trait (chosen by the highest PIP). This chosen value was defined as the trait-specific HFS for this locus, and was used for downstream trait association analysis.

HFS-trait association

For each locus, we calculated the association between trait-specific HFS and adjusted, normalized trait value by linear regression, without any covariates (this is because all selected covariates have been adjusted at the normalization step). For uniformity, we set the significance threshold at $p<5\times10^{-8}$, even if it was over-stringent for n=590,959 loci. Among significant associations, we defined an independent association as the locus with the lowest p value in the 200kb regions. As a positive control, we applied quantitative and binary GWAS with REGENIE[46], using default settings and the same British training sample. The main difference is that we used raw trait values in REGENIE, and provided the same covariates. We calculated the genomic control inflation factor, $\lambda_{GC,}$ as the median of X^2 statistics, separately for HFS association test and GWAS (only those SNPs in hapmap3[47] project were calculated). We compared the λ_{GC} between HFS and SNP by Pearson correlation analysis and paired t-test.

Fine-mapping analysis

We divided hg38 genome into 1,361 independent blocks as defined by MacDonald et al.[23], and applied SUSIE to HFS of all loci within each block, separately for each

trait (parameters: maximum number of causal signal=10, coverage=0.95). We subtracted reference HFS value for each locus prior to analysis, such that homozygous reference haplotype corresponded to HFS=0. To avoid influence of sei prediction noise, we rounded the HFS value at two decimals. Loci whose HFS had PIP>0.95 were defined as causal loci, and loci that had causal association with multiple traits were defined as pleiotropic loci. As a positive control, we applied PolyFun[9] and SbayesRC[11] on the GWAS summary statistics by REGENIE on the same training set, and extracted the reported PIP to define causal SNP.

To analyze the functional characteristics of causal loci, we first defined the sequence class of each locus by the maximum sequence class score of reference haplotype. We then tested whether each sequence class contained excess causal loci of each trait by Fisher test. For each causal locus, we also defined a "control" locus as the nearest locus that matched the p-value of this causal locus, and tested whether causal loci carried more PolyFun causal SNP than control loci by Fisher test. Furthermore, For traits whose causal loci covered >0.1% of genome-wide SNP, we applied LDSC[10] to quantify the heritability enrichment in causal and control loci, and compare their difference by jackknife method. To avoid winner's curse, we used external GWAS summary statistics for this analysis[48,49]. As an alternative method to quantify the heritability captured by causal loci, we ran multivariate linear regression in independent British test set where HFS of causal loci were independent variables and trait value were dependent variable, and calculated the R^2 and AIC. We applied the same analysis on causal SNP, and compared AIC between HFS and SNP multivariate regression.

Functional enrichment analysis

Similar to the idea of LDSC[10], we first generated a series of baseline annotation of each locus, then tested whether locus PIP was associated with functional annotations after controlling the impact of these baseline annotations. Specifically, we defined the following baseline annotations:

1) number of haplotypes, range of HFS distribution of all haplotypes (scaled by reference HFS), and 39 sequence class score of reference haplotype.

2) genomic regions of conserved base, high Phastcons score[50] in mammals, primates and vertebrate, exon, intron, untranslated regions at 3' and 5' and 200bp flanking regions of TSS. We used bedtools *intersect -f 0.1* option to annotate each locus by these annotations.

3) maximum B statistics[51], minimum allele age and $ASMC_{avg}$[52] of all variants within this locus.

Type two and three annotations were directly obtained from LDSC[10] baseline annotations. We did not include annotations related to functional genomics, since 39 sequence class score were used to capture functional genomic characteristics. Conditioned on these baseline annotations, we analyzed the enrichment of PIP in the following functional annotations:

1) Biological pathways: We downloaded all pathways from MsigDB[53] C2: canonical pathways category (including Reactome[54], PID[55], Biocarta and Wikipathway) and C6: Gene ontology[56] (biological process) category. We retained

only pathways with >5 genes and <500 genes. We generated a gene × pathway binary matrix and applied hierarchical clustering so that similar pathways were placed close to each other. We sequentially compared adjacent pathways, and removed the smaller one if the fraction of overlap > 30%. A total of 3,219 pathways were retained. We then linked each locus to these pathways by cS2G[24] strategy. Specifically, a locus L would be annotated as 1 for pathway P only if L contained a SNP that was link to P with cS2G score>0.5.

2) Tissue-specific chromatin activity: We downloaded chromHMM[57] chromatin state annotation for 833 samples from epimap[40], and grouped them according to developmental stages and second-level tissue types. For each group, all chromosomal regions annotated as "transcription start site (TSS), transcription region (TX), enhancer, promoter" in at least half of the samples were marked as active regions. We used bedtools *intersect -f 0.1* option to annotate whether each locus was active in each tissue.

3) cell type-specific open chromatin regions: we downloaded scATAC-seq peak data from Zhang et al.[58], and annotated each locus by bedtools *intersect -f 0.1* option.

We applied multivariate linear regression of PIP against baseline annotations + one of the functional annotations. Regression coefficient > 0 and Bonferroni-adjusted regression p value<0.05 were used as significance threshold. From the final results, we manually removed those pathways and cell types that reached significance threshold in more than half of the traits, since these pathways likely reflected unrecognized confounders.

Polygenic prediction

We used the posterior effect size estimated by SUSIE as weights, and calculated the weighted sum of HFS as the polygenic risk score of each trait, and calculated R^2 in independent British test sample with simple linear regression. As a positive control, we applied LDAK-BOLT[30] algorithm on the SNP array data (about seven hundred thousand variants) with tenfold cross-validation and max iteration=200 in the same training sample, and calculated SNP-based PRS with the output SNP weights. Normalized trait values were analyzed, without any covariates provided. Array data were filtered by Plink with option --geno 0.1 --hwe 1e-6 --mac 100 --maf 0.01 --mind 0.1.

To train an FCNN model that predict height, we first extracted all HFS with PIP>0.4, subtracted reference HFS value, and scaled to -1 ~ 1 range. We applied Bayesian optimization to choose number of hidden layers (2 to 3), number of nodes per layers (50 to 150), learning rate (10^{-5} to 10^{-1}) and alpha (10^{-4} to 0.9). We set the default activation function as linear, although we tried other activation functions after parameter tuning. We trained FCNN with MLPregressor function. Evaluation metric was defined as Root Mean Square Error for all FCNN analysis.

For non-British European (NBE), South Asian (SAS), East Asian (EA) and African (AFR) participants in UK Biobank, we randomly selected half as tuning sample and half as test sample. In the tuning sample, we applied the trained FCNN model and obtained output of the last hidden layers (number of nodes=133). Then, we

calculated SNP-based PRS and used these 134 features in an elastic net by glmnet R package. This is equivalent to fix the hidden layers of FCNN, add SNP-based PRS to the output layer and train the output layer in the tuning sample, which is a classic transfer learning framework. We used five-fold cross validation in elastic net, and the search grid was set at 0.01 to 1 for alpha and 0.001 to 0.5 for lambda. The R^2 of HFS_{FCNN}+LDAK and LDAK along were calculated in the ancestry-specific test set. Additionally, we applied PolyFun-pred[33] and SbayesRC[11] to the summary statistics of height (calculated by REGENIE in the same training sample), and integrated their effect size with LDAK weight in the tuning sample using Polypred[33] method. PRS for LDAK, LDAK+PolyFun and LDAK+SbayesRC were calculated by plink *score* option, excluding variants with INFO<0.8, Hardy-Weinberg $p<10^{-6}$, allele count<2 or missing rate >10% in the target test set.

Statistical analysis

All p-values were two-sided and adjusted by Bonferroni unless otherwise specified. For group comparison, we used Fisher test for count data and paired t-test for continuous data. For R^2 of PRS comparison, we applied r2redux[59] R package to estimate 95% confidence interval and its p-value for the difference of R^2.

Declaration

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

This work was based on UK Biobank data from project 84436. Code for this analysis is available at https://github.com/WeiCSong/HFS

Acknowledgement

Not applicable.

Funding

This work was supported by grants from the 2030 Science and Technology Innovation Key Program of Ministry of Science and Technology of China (No. 2022ZD020910001), the National Natural Science Foundation of China (No. 81971292, 82150610506) and the Natural Science Foundation of Shanghai (No. 21ZR1428600), the Medical-Engineering Cross Foundation of Shanghai Jiao Tong University (No. YG2022ZD026).

Authors' contributions

W.S designed the study, collected and analyzed the data and wrote the manuscript. G.N.L and Y.S supervised the study. All authors read, revise and approved the

manuscript.

Competing interests

The authors declare that they have no competing interests.

Reference

1. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017 Jul 6;101(1):5–22.
2. Li X, Li Z, Zhou H, Gaynor SM, Liu Y, Chen H, et al. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. Nat Genet. 2020 Aug 24;52(9):969–83.
3. Null M, Dupuis J, Sheinidashtegol P, Layer RM, Gignoux CR, Hendricks AE. RAREsim: A simulation method for very rare genetic variants. Am J Hum Genet. 2022 Mar;0(0).
4. Zhou W, Bi W, Zhao Z, Dey KK, Jagadeesh KA, Karczewski KJ, et al. SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests. Nat Genet. 2022 Sep 22;54(10):1466–9.
5. Zeng J, Xue A, Jiang L, Lloyd-Jones LR, Wu Y, Wang H, et al. Widespread signatures of natural selection across human complex traits and functional genomic categories. Nat Commun. 2021 Sep 1;12(1):752527.
6. Nowbandegani PS, Wohns AW, Ballard JL, Lander ES, Bloemendal A, Neale BM, et al. Extremely sparse models of linkage disequilibrium in ancestrally diverse association studies. bioRxiv. 2022 Sep 8;2022.09.06.506858.
7. Watanabe K, Stringer S, Frei O, Umićević Mirkov M, de Leeuw C, Polderman TJC, et al. A global overview of pleiotropy and genetic architecture in complex traits. Nat Genet. 2019 Sep 1;51(9):1339–48.
8. Iotchkova V, Ritchie GRS, Geihs M, Morganella S, Min JL, Walter K, et al. GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. Nat Genet. 2019 Feb 1;51(2):343–53.
9. Weissbrod O, Hormozdiari F, Benner C, Cui R, Ulirsch J, Gazal S, et al. Functionally informed fine-mapping and polygenic localization of complex trait heritability. Nat Genet. 2020 Dec 1;52(12):1355–63.
10. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015 Nov 28;47(11):1228–35.
11. Zheng Z, Liu S, Sidorenko J, Yengo L, Turley P, Ani A, et al. Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries. bioRxiv. 2022 Oct 14;2022.10.12.510418.
12. Grotzinger AD, de la Fuente J, Davies G, Nivard MG, Tucker-Drob EM. Transcriptome-wide and stratified genomic structural equation modeling

identify neurobiological pathways shared across diverse cognitive traits. Nat Commun 2022 131. 2022 Oct 21;13(1):1–15.
13. Liang Y, Pividori M, Manichaikul A, Palmer AA, Cox NJ, Wheeler HE, et al. Polygenic transcriptome risk scores (PTRS) can improve portability of polygenic risk scores across ancestries. Genome Biol. 2022 Dec 1;23(1):1–18.
14. Hu X, Qiao D, Kim W, Moll M, Balte PP, Lange LA, et al. Polygenic transcriptome risk scores for COPD and lung function improve cross-ethnic portability of prediction in the NHLBI TOPMed program. Am J Hum Genet. 2022 Apr;0(0).
15. Aguet F, Barbeira AN, Bonazzola R, Brown A, Castel SE, Jo B, et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science (80-). 2020 Sep 1;369(6509):1318–30.
16. Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun. 2018 Dec;9(1):1–20.
17. Chen KM, Wong AK, Troyanskaya OG, Zhou J. A sequence-based global map of regulatory activity for deciphering human genetics. Nat Genet. 2022 Jul 11;54(7):940–9.
18. Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021 Oct 4;18(10):1196.
19. Kelley DR. Cross-species regulatory sequence activity prediction. PLoS Comput Biol. 2020 Jul 1;16(7).
20. Yan J, Qiu Y, Ribeiro dos Santos AM, Yin Y, Li YE, Vinckier N, et al. Systematic analysis of binding of transcription factors to noncoding variants. Nature. 2021 Jan 27;591(7848):147–51.
21. Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat Genet. 2018 Aug 1;50(8):1171–9.
22. Wang G, Sarkar A, Carbonetto P, Stephens M. A simple new approach to variable selection in regression, with application to genetic fine mapping. J R Stat Soc Ser B (Statistical Methodol. 2020 Dec 1;82(5):1273–300.
23. MacDonald JW, Harrison T, Bammler TK, Mancuso N, Lindström S. An updated map of GRCh38 linkage disequilibrium blocks based on European ancestry data. bioRxiv. 2022 Mar 7;(December 2018):2022.03.04.483057.
24. Gazal S, Weissbrod O, Hormozdiari F, Dey KK, Nasser J, Jagadeesh KA, et al. Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. Nat Genet. 2022 Jun 6;54(6):827–36.
25. Finucane HK, Reshef YA, Anttila V, Slowikowski K, Gusev A, Byrnes A, et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat Genet. 2018 Apr 9;50(4):621–9.
26. Morales MAG, Montero□vargas JM, Vizuet□de□rueda JC, Teran LM. New Insights into the Role of PD-1 and Its Ligands in Allergic Disease. Int J Mol

Sci. 2021 Nov 1;22(21):11898.
27. Kong X, Ma L, Chen E, Shaw CA, Edelstein LC. Identification of the Regulatory Elements and Target Genes of Megakaryopoietic Transcription Factor MEF2C. Thromb Haemost. 2019;119(5):716.
28. Sekar A, Bialas AR, Rivera H de, Davis A, Hammond TR, Kamitaki N, et al. Schizophrenia risk from complex variation of complement component 4. Nature. 2016 Feb 11;530(7589):177.
29. Nawijn MC, Piavaux BJA, Jeurink P V., Gras R, Reinders MA, Stearns T, et al. Identification of the Mhc Region as an Asthma Susceptibility Locus in Recombinant Congenic Mice. Am J Respir Cell Mol Biol. 2011 Aug 1;45(2):295.
30. Zhang Q, Privé F, Vilhjálmsson B, Speed D. Improved genetic prediction of complex traits from individual-level data or summary statistics. Nat Commun. 2021 Jul 7;12(1):1–9.
31. Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AAE, Lee SH, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat Genet. 2015 Aug 31;47(10):1114–20.
32. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010 Jun 20;42(7):565–9.
33. Weissbrod O, Kanai M, Shi H, Gazal S, Peyrot WJ, Khera A V., et al. Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. Nat Genet. 2022 Apr 7;54(4):450–8.
34. Helland IB, Smith L, Saarem K, Saugstad OD, Drevon CA. Maternal Supplementation With Very-Long-Chain n-3 Fatty Acids During Pregnancy and Lactation Augments Children's IQ at 4 Years of Age. Pediatrics. 2003 Jan 1;111(1):e39–44.
35. Park CY, Zhou J, Wong AK, Chen KM, Theesfeld CL, Darnell RB, et al. Genome-wide landscape of RNA-binding protein target site dysregulation reveals a major impact on psychiatric disorder risk. Nat Genet. 2021 Feb 1;53(2):166.
36. Zhou J. Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. Nat Genet. 2022 May 12;54(5):725–34.
37. Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam HJ, et al. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nat Commun. 2020 Nov 20;11(1):1–13.
38. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nat 2021 5967873. 2021 Jul 15;596(7873):583–9.
39. Liu Z, Pan W, Li W, Zhen X, Liang J, Cai W, et al. Evaluation of the Effectiveness of Derived Features of AlphaFold2 on Single-Sequence Protein Binding Site Prediction. Biology (Basel). 2022 Oct 1;11(10):1454.
40. Boix CA, James BT, Park YP, Meuleman W, Kellis M. Regulatory genomic

circuitry of human disease loci by integrative epigenomics. Nature. 2021 Feb 3;590(7845):300–7.
41. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841–2.
42. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
43. Zhao H, Sun Z, Wang J, Huang H, Kocher JP, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics. 2014 Apr 1;30(7):1006–7.
44. Delaneau O, Zagury JF, Robinson MR, Marchini JL, Dermitzakis ET. Accurate, scalable and integrative haplotype estimation. Nat Commun 2019 101. 2019 Nov 28;10(1):1–10.
45. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021 Jan 29;10(2):1–4.
46. Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, et al. Computationally efficient whole-genome regression for quantitative and binary traits. Nat Genet. 2021 May 20;53(7):1097–103.
47. Altshuler DM, Gibbs RA, Peltonen L, Schaffner SF, Yu F, Dermitzakis E, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010 Sep 2;467(7311):52–8.
48. Yengo L, Vedantam S, Marouli E, Sidorenko J, Bartell E, Sakaue S, et al. A saturated map of common genetic variants associated with human height. Nature. 2022 Oct 12;610(7933):704–12.
49. Mikaelsdottir E, Thorleifsson G, Stefansdottir L, Halldorsson G, Sigurdsson JK, Lund SH, et al. Genetic variants associated with platelet count are predictive of human disease and physiological markers. Commun Biol. 2021 Dec 1;4(1).
50. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005 Aug 1;15(8):1034–50.
51. McVicker G, Gordon D, Davis C, Green P. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 2009 May;5(5).
52. Palamara PF, Terhorst J, Song YS, Price AL. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. Nat Genet. 2018 Sep 1;50(9):1311–7.
53. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545–50.
54. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2018 Jan 1;46(D1):D649–55.
55. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, et al. PID: the Pathway Interaction Database. Nucleic Acids Res. 2009;37(Database

issue):D674.
56. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000 May;25(1):25–9.
57. Ernst J, Kellis M. ChromHMM: Automating chromatin-state discovery and characterization. Vol. 9, Nature Methods. 2012. p. 215–6.
58. Zhang K, Hocker JD, Miller M, Wang A, Preissl S, Correspondence BR, et al. A single-cell atlas of chromatin accessibility in the human genome. Cell. 2021;184:1–17.
59. Momin MM, Lee S, Wray NR, Lee SH. Significance tests for R2 of out-of-sample prediction using polygenic scores. Am J Hum Genet. 2023 Jan;0(0).

Figure Legends

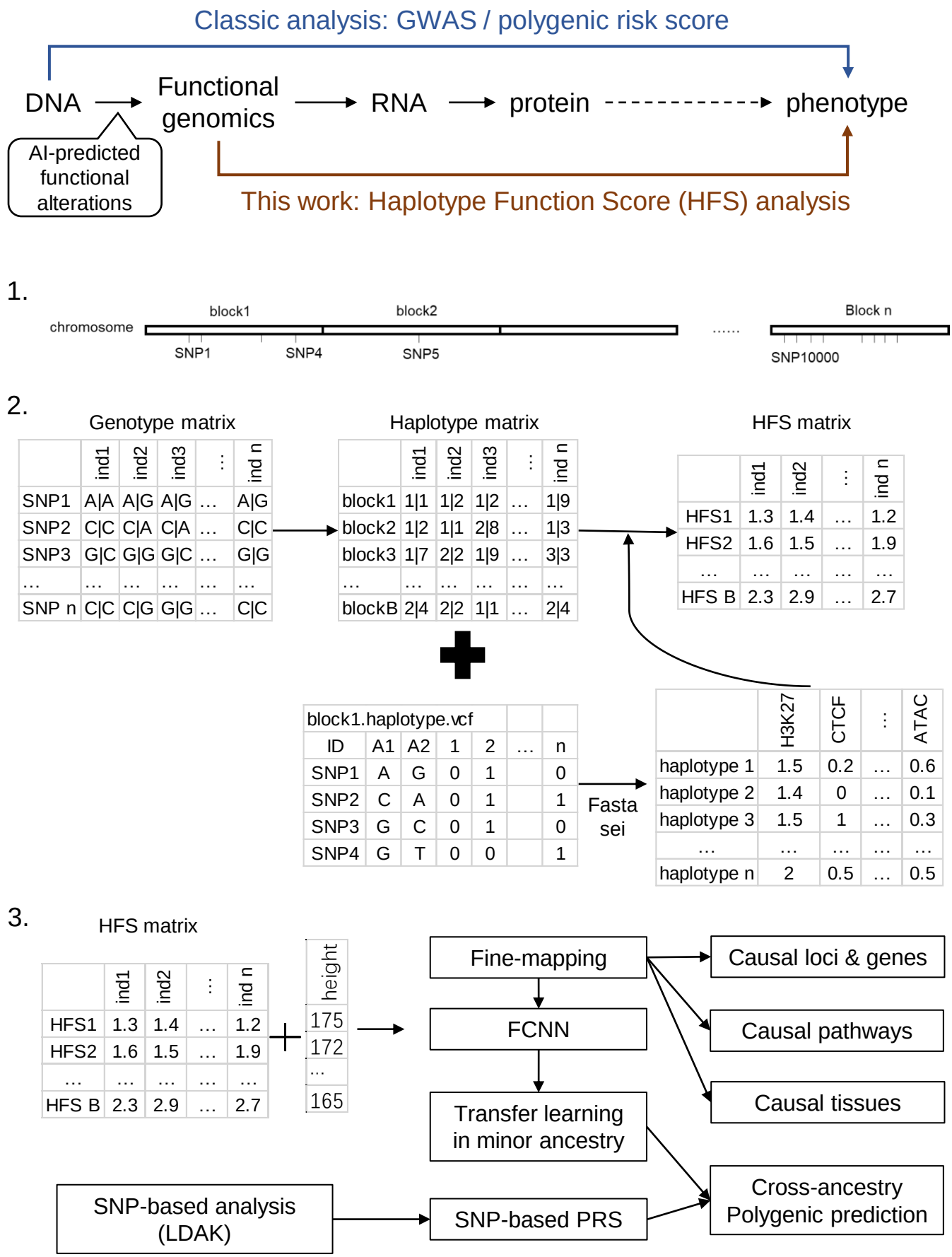
Figure 1 Flowchart of the study. Ind: individual.

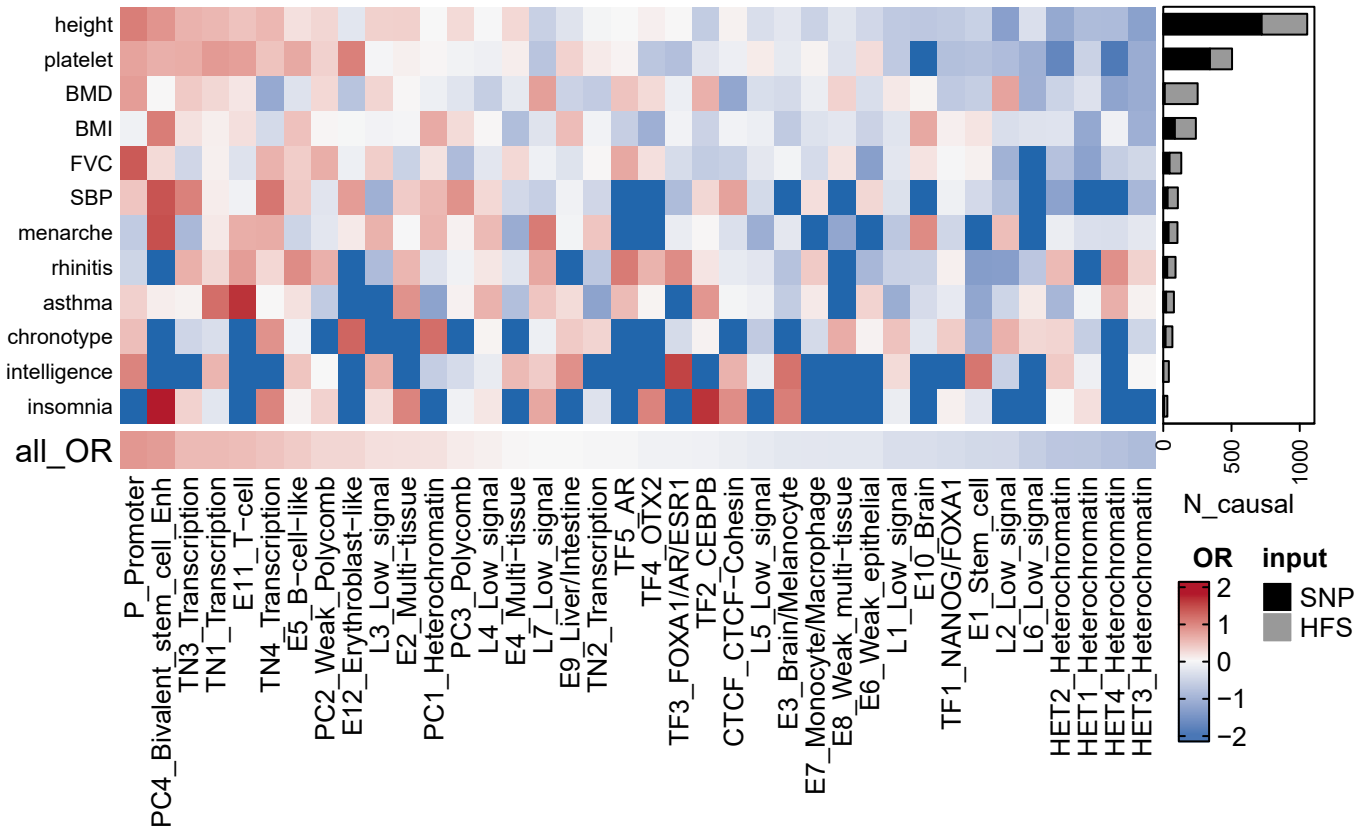
Figure 2 Fine-mapping result summary. Grey bar plots indicated the number of loci with posterior inclusion probability (PIP)>0.95 in HFS+SUSIE (causal loci). Black bar plots indicated number of SNP with PIP>0.95 in PolyFun or SbayesRC analysis (the larger number was shown). Each grid of heatmap showed the Odds ratio of each sequence class loci being causal loci for each trait. “All_OR” indicated Odds ratio for pooling all traits together. Enh: enhancer. TF: transcription factor binding site.

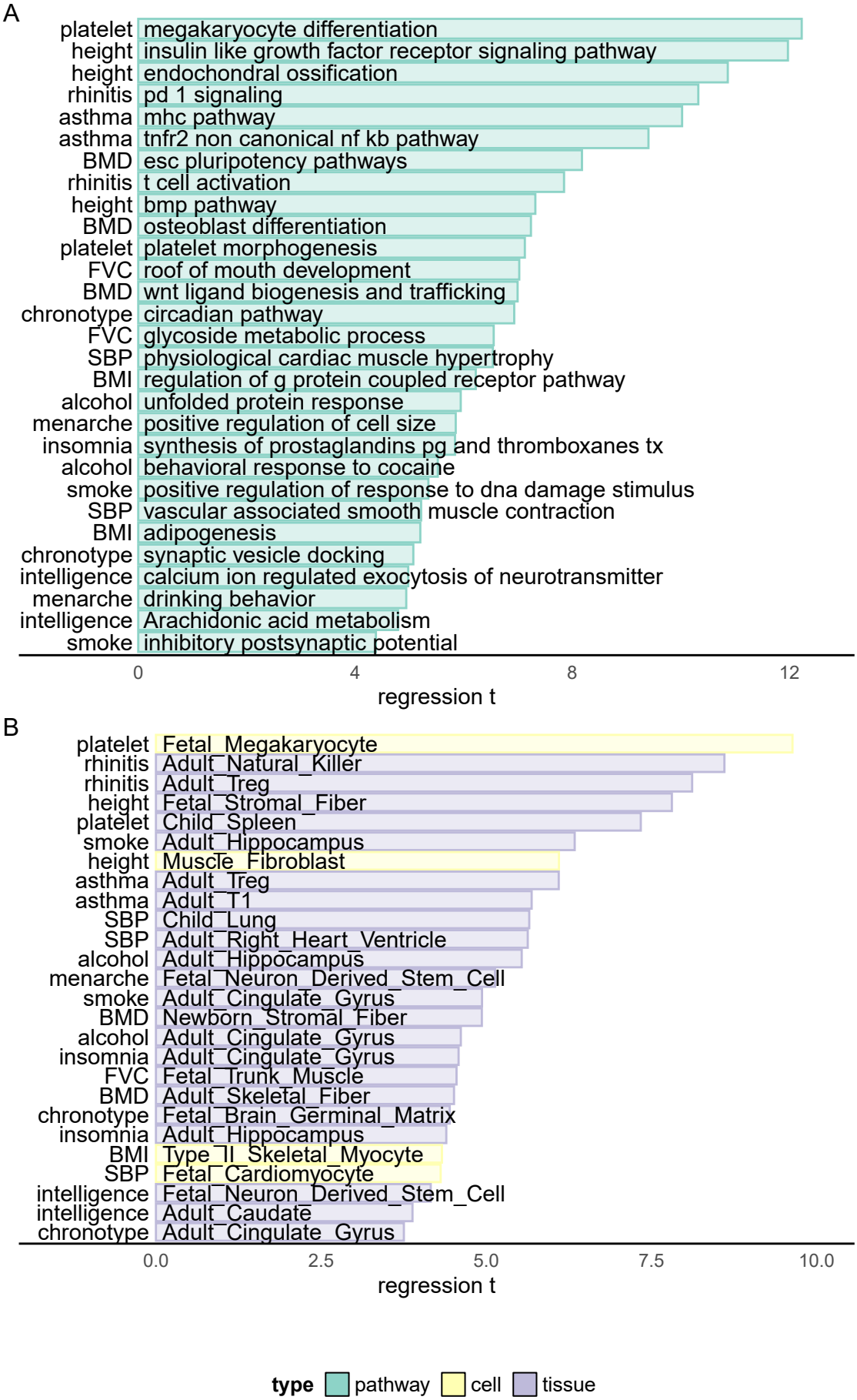
Figure 3 Biological enrichment analysis based on HFS fine-mapping. X-axis indicated t statistics of the analyzed term in a multivariate linear regression (Method). Cell: single-cell ATAC peak for 222 cell types from Zhang et al[58]. Tissue: active chromatin regions of 222 tissues from epimap[40]. For each trait, we showed the most significant term plus one or two terms with high biological interpretation that also passed significance threshold. Full enrichment result was shown in Table S7 and S8.

Figure 4 HFS linked trait to causal genes. A: Target genes of causal loci identified by HFS+SUSIE for platelet count. Only genes that showed functional convergence were shown. B: Regional plot for RBBP5. HFS: loci PIP calculated by HFS+SUSIE. SNP: SNP PIP calculated by PolyFun. cCRE: credible cis-regulation elements. C: Regional plot of MHC region for asthma. Thickened curve linked highlighted causal loci to its target genes predicted by cS2G[24].

Figure 5 HFS-based polygenic prediction. A: Prediction R^2 of HFS-based polygenic risk score (PRS) using different threshold of PIP. allSNP: SNP-based PRS calculated by LDAK-BOLT[30]. n: number of features included in the corresponding PRS. B: Prediction R^2 of HFS-based fully-connected neural network (FCNN) and HFS+SUSIE. C: Prediction R^2 of different tools in non-British European (NBE), South Asian (SAS), East Asian (EAS) and African (AFR) groups in UK Biobank.


Classic analysis: GWAS / polygenic risk score
DNA
Functional genomics
RNA
protein
phenotype
AI-predicted functional alterations
This work: Haplotype Function Score (HFS) analysis
1.
chromosome
block1
block2
Block n
SNP1
SNP4
SNP5
SNP10000
2.
Genotype matrix
Haplotype matrix
HFS matrix
block1.haplotype.vcf
Fasta sei
3.
HFS matrix
height
Fine-mapping
FCNN
Transfer learning in minor ancestry
SNP-based analysis (LDAK)
SNP-based PRS
Causal loci & genes
Causal pathways
Causal tissues
Cross-ancestry Polygenic prediction

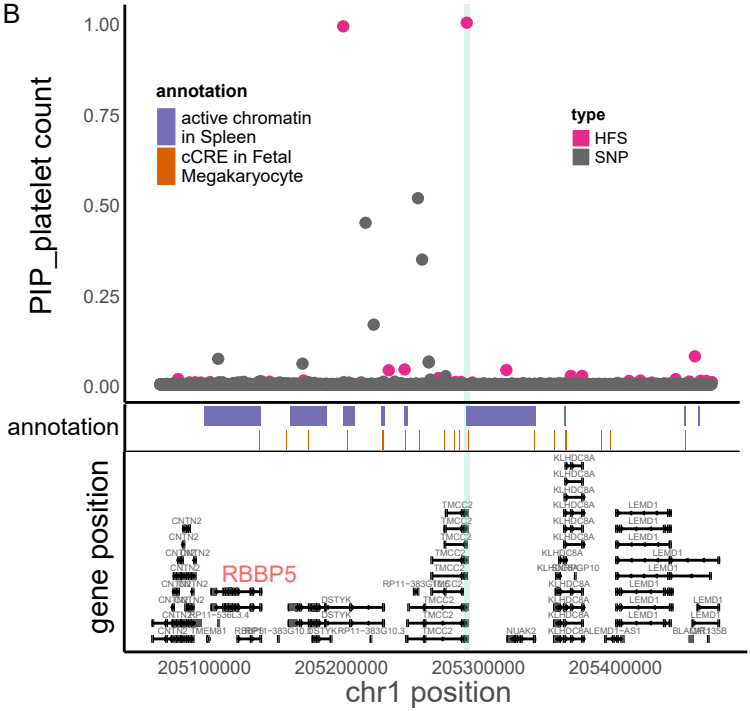

height
platelet
BMD
BMI
FVC
SBP
menarche
rhinitis
asthma
chronotype
intelligence
insomnia
all_OR
P_Promoter
PC4_Bivalent_stem_cell_Enh
TN3_Transcription
TN1_Transcription
E11_T−cell
TN4_Transcription
E5_B−cell−like
PC2_Weak_Polycomb
E12_Erythroblast−like
L3_Low_signal
E2_Multi−tissue
PC1_Heterochromatin
PC3_Polycomb
L4_Low_signal
E4_Multi−tissue
L7_Low_signal
E9_Liver/Intestine
TN2_Transcription
TF5_AR
TF4_OTX2
TF3_FOXA1/AR/ESR1
TF2_CEBPB
CTCF_CTCF−Cohesin
L5_Low_signal
E3_Brain/Melanocyte
E7_Monocyte/Macrophage
E8_Weak_multi−tissue
E6_Weak_epithelial
L1_Low_signal
E10_Brain
TF1_NANOG/FOXA1
E1_Stem_cell
L2_Low_signal
L6_Low_signal
HET2_Heterochromatin
HET1_Heterochromatin
HET4_Heterochromatin
HET3_Heterochromatin
0
500
1000
N_causal
OR
2
1
0
−1
−2
input
SNP
HFS


A
platelet megakaryocyte differentiation
height insulin like growth factor receptor signaling pathway
height endochondral ossification
rhinitis pd 1 signaling
asthma mhc pathway
asthma tnfr2 non canonical nf kb pathway
BMD esc pluripotency pathways
rhinitis t cell activation
height bmp pathway
BMD osteoblast differentiation
platelet platelet morphogenesis
FVC roof of mouth development
BMD wnt ligand biogenesis and trafficking
chronotype circadian pathway
FVC glycoside metabolic process
SBP physiological cardiac muscle hypertrophy
BMI regulation of g protein coupled receptor pathway
alcohol unfolded protein response
menarche positive regulation of cell size
insomnia synthesis of prostaglandins pg and thromboxanes tx
alcohol behavioral response to cocaine
smoke positive regulation of response to dna damage stimulus
SBP vascular associated smooth muscle contraction
BMI adipogenesis
chronotype synaptic vesicle docking
intelligence calcium ion regulated exocytosis of neurotransmitter
menarche drinking behavior
intelligence Arachidonic acid metabolism
smoke inhibitory postsynaptic potential
0
4
8
12
regression t
B
platelet Fetal_Megakaryocyte
rhinitis Adult_Natural_Killer
rhinitis Adult_Treg
height Fetal_Stromal_Fiber
platelet Child_Spleen
smoke Adult_Hippocampus
height Muscle_Fibroblast
asthma Adult_Treg
asthma Adult_T1
SBP Child_Lung
SBP Adult_Right_Heart_Ventricle
alcohol Adult_Hippocampus
menarche Fetal_Neuron_Derived_Stem_Cell
smoke Adult_Cingulate_Gyrus
BMD Newborn_Stromal_Fiber
alcohol Adult_Cingulate_Gyrus
insomnia Adult_Cingulate_Gyrus
FVC Fetal_Trunk_Muscle
BMD Adult_Skeletal_Fiber
chronotype Fetal_Brain_Germinal_Matrix
insomnia Adult_Hippocampus
BMI Type_II_Skeletal_Myocyte
SBP Fetal_Cardiomyocyte
intelligence Fetal_Neuron_Derived_Stem_Cell
intelligence Adult_Caudate
chronotype Adult_Cingulate_Gyrus
0.0
2.5
5.0
7.5
10.0
regression t
type
pathway
cell
tissue

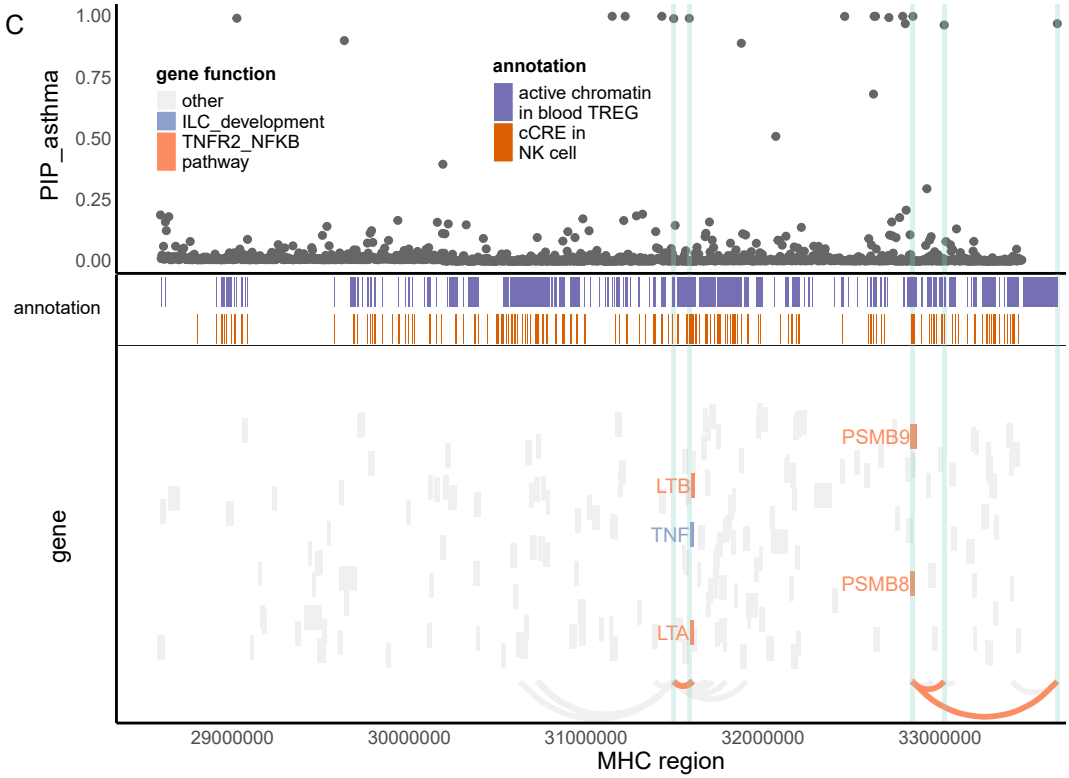
A

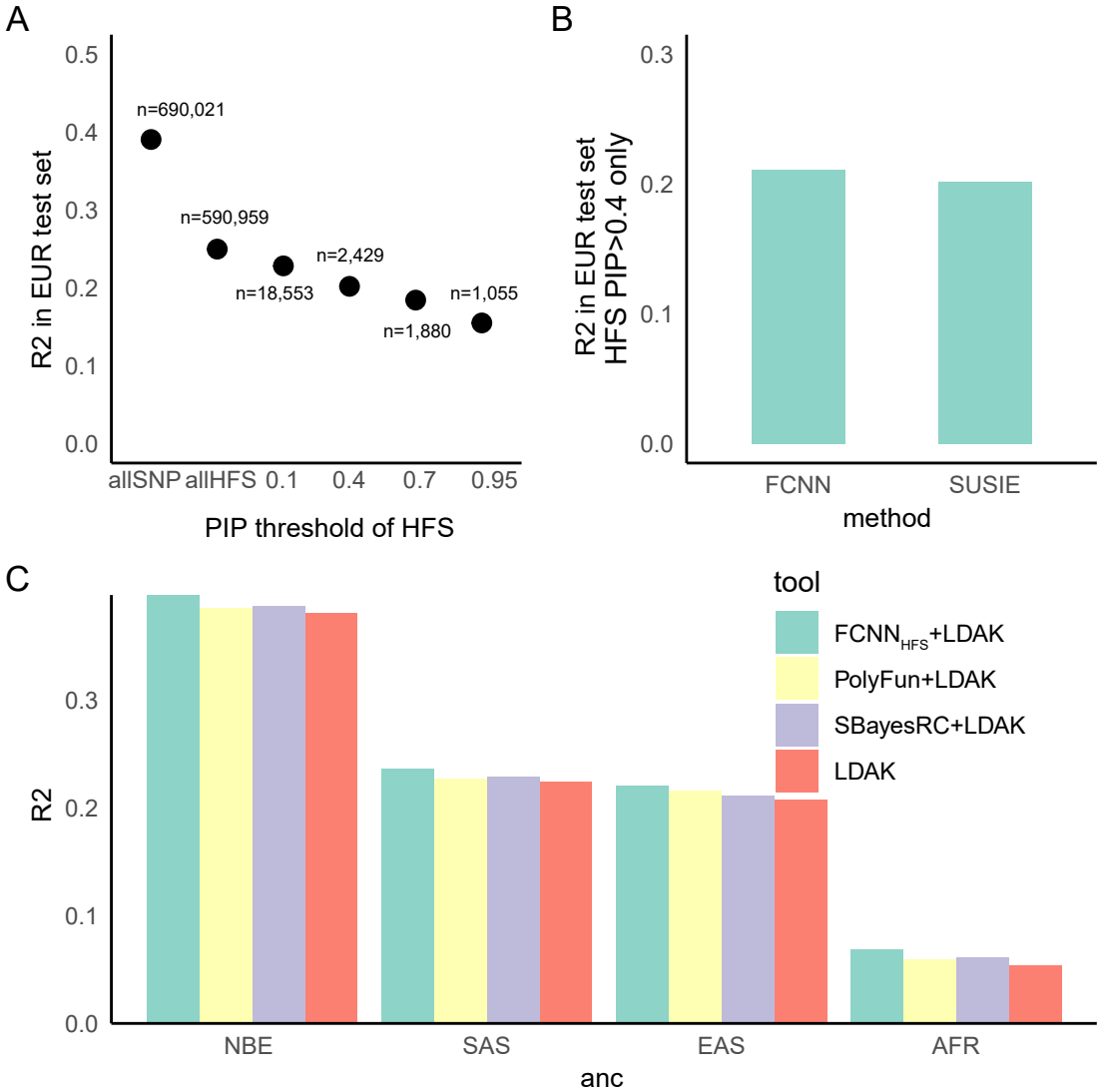
HFS position	gene	HFS_based PIP	MK Differentiation	Platelet Morphogenesis	Fetal Megakaryocyte	Child Spleen
3q27.1	THPO	1	■			■
1q32.1	RBBP5	1	■		■	■
12q24.12	SH2B3	1	■			■
12q13.13	NFE2	1	■		■	
5q14.3	MEF2C	1	■	■		■
3q27.1	THPO	1	■			■
16q24.2	ZFPM1	1	■	■	■	■
11q24.3	FLI1	1	■			■
16q24.2	ZFPM1	1	■	■	■	■
3q27.1	THPO	1	■			■
11q24.3	FLI1	1	■		■	■
5q14.3	MEF2C	1	■	■		■
12q13.13	NFE2	1	■			■
12q24.12	SH2B3	1	■		■	■
12q24.12	SH2B3	1	■			■
21q22.12	RUNX1	1	■			■
5q14.3	MEF2C	1	■	■		■
3q27.1	THPO	1	■		■	■
1p33	STIL	0.9	■	■	■	
15q14	THBS1	0.9	■			■
11q24.3	FLI1	0.8	■		■	■
11q24.3	FLI1	0.8	■		■	■
12q13.13	NFE2	0.8	■	■	■	■
14q24.1	DCAF5	0.8		■	■	■
3q21.3	GATA2	0.7	■		■	■

B



C




A
R2 in EUR test set
0.5
0.4
0.3
0.2
0.1
0.0
n=690,021
n=590,959
n=18,553
n=2,429
n=1,880
n=1,055
allSNP
allHFS
0.1
0.4
0.7
0.95
PIP threshold of HFS
B
R2 in EUR test set
HFS PIP>0.4 only
0.3
0.2
0.1
0.0
FCNN
SUSIE
method
C
R2
0.3
0.2
0.1
0.0
NBE
SAS
EAS
AFR
anc
tool
FCNN$_{HFS}$+LDAK
PolyFun+LDAK
SBayesRC+LDAK
LDAK